VIMALA COLLEGE (AUTONOMOUS)

(NAAC Re-accredited (3rd Cycle): A Grade, CGPA-3.50)

(Affiliated to University of Calicut)

POST GRADUATE DIPLOMA

IN

DATA SCIENCE

(CHOICE BASED CREDIT AND SEMESTER SYSTEM)

UNDER THE

FACULTY OF SCIENCE

SYLLABUS

(FOR THE STUDENTS ADMITTED FROM THE ACADEMIC YEAR 2018 – 19 ONWARDS)

VIMALA COLLEGE

ENGINEERING COLLEGE P O, THRISSUR

KERALA-680009

INDIA

REGULATIONS FOR THE DEGREE OF
Post Graduate Diploma in DATA SCIENCE

1 PROGRAMME OBJECTIVES

The course of the PG Diploma in Data Science programme is designed with the following objectives:

a) To equip students to take up challenging research oriented responsibilities and courses for their higher studies/profession.

b) To train and equip the students to meet the industry standards in field of Data Science.

c) To equip with the knowledge in business intelligence and business data analysis.

d) To develop quantitative, analytical and visualization skills needed to solve a given problem.

e) To understand the world of humongous data and how to transform it to knowledge

f) To develop in depth theoretical knowledge about the various statistical methods

g) To have practical and experiential learning on various analytical models

2 GENERAL PROGRAMME STRUCTURE

Duration: The duration of the PG Diploma in Data Science programme shall be 2 semesters distributed over a period of one academic year. The odd semester shall be from June to November and the even Semester shall be from December to May. Each semester shall have 105 working days inclusive of all examinations.

Courses: The programme includes Core courses which are theory and practical oriented. There is a Project Work which is to be undertaken by all students. No course shall have more than 4 credits. For project work and General Viva-Voce, the maximum credits shall be 8

Attendance: A student shall be permitted to appear for the semester examination, only if (s)he secures not less than 75% attendance in each semester. Attendance shall be maintained by the department. Condonation of shortage of attendance to a maximum of 10 days in a semester during the whole period of the PG Diploma in Data Science programme may be granted. Benefits of attendance may be granted to students who attend the approved activities of college/university with prior concurrence of the head of the institution. Participation in such activities may be treated as presence in lieu of their absence on production of participation / attendance certificate in curricular / extracurricular activities. It should be limited to 10 days in a semester.

If a student registered in first semester of the PG Diploma in Data Science programme is continuously absent from the classes for more than 14 working days at the beginning of the semester without informing the authorities the matter shall immediately be brought to the notice of the Principal of the college. The names of such students shall be removed from the rolls.

Admission to repeat courses should be within the sanctioned strength. However if more candidates are there, the candidates who have suffered serious health problems, on production of a medical certificate issued by a physician not below the rank of a Civil Surgeon in Government service, may be permitted to repeat the course. The number of such candidates should not exceed two.

Project: Every student of the PG Diploma in Data Science programme shall have to work on a project of not less than 8 credits under the supervision of a faculty member as per the curriculum.

3 ADMISSION

The admission to all programmes will be as per the rules and regulations of the Institution. The eligibility criteria for admission shall be as announced by the Institution from time to time. However the minimum eligibility of the course is a UG degree in Computer Science/ IT/ BCA or UG degree from any discipline with Statistics as core/complementary. The maximum intake of the course is 12.

Separate rank lists shall be drawn up for reserved seats as per the existing rules. The college shall make available to all students admitted a prospectus listing all the courses offered including electives in various departments during a particular semester. The information provided shall contain title of the course and credits of the course.

There shall be a uniform calendar prepared by the Institution for the registration, conduct /schedule of the courses, examinations and publication of results. The Institution shall ensure that the calendar is strictly followed.

4 REGISTRATION

A student shall be permitted to register for the programme at the time of admission. A student shall be permitted to register for the examination also. If registration for examination is not possible owing to shortage of attendance beyond condonation limit, the student shall not be permitted to move to the next semester.

5 COURSE EVALUATION

The evaluation scheme for each course shall contain two parts: (a) internal evaluation and (b) external evaluation. 20% weight shall be given to internal evaluation and the remaining 80% to external evaluation. Therefore the ratio of weight between internal and external is 1:4. Both internal and external evaluation shall be carried out using Direct grading system.

INTERNAL EVALUATION

The internal evaluation shall be based on predetermined transparent system involving periodic written tests, assignments, seminars and attendance in respect of theory courses and based on lab tests, lab skill/records/viva and attendance in respect of practical courses.

THEORY PAPERS

The weightage assigned to various components for internal evaluation for theory papers is as shown below.

Components	Weightage
Test papers with at least 25% questions based on problems or programs (minimum two)	2
Assignments (minimum two) such as homework, problem solving, group discussions, quiz, literature survey, term-project, software exercises, etc.	1
Regularity in the class	1
Seminar	1
Total	5

To ensure transparency of the evaluation process, the internal assessment grade awarded to the students in each course in a semester shall be published on the notice board at least one week before the commencement of external examination. There shall not be any chance for improvement for internal grade.

The course teacher shall maintain the academic record of each student registered for the course, which shall be forwarded to the Institution, through the college Principal.

PRACTICAL PAPERS

The mark distribution to award internal continuous assessment marks for practical course should be as follows:

Components	Weightage
Rough record for each experiment	1
Performance in the laboratory – coding, results	1
Fair Record	1
Regularity	1
End-semester test	1
Total	5

Note:

1. All students should have a rough record (observation note book) in which they write all the works to be carried out in the lab prior to his/her entering the lab. (S)he may also note down the i/p and o/p that (s)he gives for program verification in the observation note book (rough record).

2. All lab works should be neatly recorded in a Laboratory Record Book (Fair Record) in written form. However program results can be pasted in the left hand side of the fare record.

3. Chairperson, Board of Examination (PG) has to prepare the modalities of the practical papers (list of experiments to be done, number of minimum experiments required in the practical record etc) and distributed to all departments concerned, at the beginning of each semester itself. Model lists of experiments are provided with the syllabus for each practical session.

4. No candidate will be permitted to attend the end-semester test unless he/she produces certified record of the laboratory.

5. Full credit for regularity in the class can be given only if the candidate has secured minimum 90% attendance in the course. Attendance evaluation for each course is as follows:

Percentage of Attendance	Weightage
90% and above	4
85 to 89.9%	3
80 to 84.9%	2
75 to 79.9%	1
Below 75 %	0

EVALUATION COMMITTEE (EC)

For the evaluation of the Project Work, an evaluation committee is to be constituted. One faculty is to be designated as the Course Coordinator for these courses. Committee is to be constituted by the head of the department (HOD) and (s)he shall be the Chairperson of the committee. In addition to the HOD, the Course Coordinator and at least three faculty members can be designated as the members of the committee. In case HOD is unable to represent himself/herself in the committee, (s)he can nominate a faculty in lieu for him/her as a member and the chairperson of the committee. In addition to this, faculty guiding a particular student will also be a member of the committee. At least one member of the committee should be a lady, if lady faculties are available in the department concerned. The Coordinator has to set the schedule for presentation and submission of the reports. While calculating the final score, 20% weight is to be given for the scores awarded by the guide to the student and the rest 80% weight is to be given for the average of the scores awarded to the student by remaining committee members.

PROJECT WORK

The project work will be assessed on the following criteria:

Component	Grade Points
Relevance of the Topic, Statement of Objectives, Methodology	15
Quality of Literature Survey/Product Review	15
Quality of Analysis Phase	10
Quality of Design Phase	10
Quality of Implementation/Simulation	10
Quality of Testing/Result Analysis	10
Quality of Contributions	10
Identification of Future Work	15
Quality of Project Report	50
Publications/Presentations out of the Project Work*	10
Quality of Presentation	15
Demonstration of the Project Work	15
General Viva Voce	15
Total	200

Grade is calculated by dividing total number of points obtained by a student by 50

*In case at least one student of the batch has a publication/presentation out of his/her project work in a workshop/conference/journal/IT fest etc, this score is to be awarded for

the student; no other students will deserve score for this component! If none of the students in the batch could make such an edge, then the score for this component is to be added with the component "Identification of Future Work".

The Evaluation Committee can decide about the components for monthly evaluation from the above list. See Appendix B for a sample evaluation.

EXTERNAL EVALUATION

- End semester examinations for theory and practical courses will be conducted by the Institution. For practical courses, end semester examinations will be conducted in every semesters.
- The external examinations in theory and practical courses are to be conducted by the Institution with question papers set by external experts.
- External project evaluation shall be conducted at the end of the second semester.
- The evaluation of the answer scripts shall be done by examiners based on a well - defined scheme of valuation. The external evaluation shall be done immediately after the examination preferably in a Centralized Valuation Camp.
- Practical examinations will be conducted by one internal and one external examiners, project evaluation shall be conducted by two external examiners.
- For Project Work, if the performance of the student is below the expected benchmark (E grade), student will be given a chance to reappear within six weeks (from the date of evaluation) to present the work again, after incorporating the changes suggested by the examiners. Examiners have to submit their suggestions in writing to Chairperson, Board of Examinations PG and the concerned HOD on the day of examination itself. HOD has to convey the matter ASAP to the students concerned. The Chairperson, Board of Examinations has to inform the concerned HOD about the schedule for resubmission and revised evaluation within seven days of the date of evaluation. While submitting the revised report, the student has to produce a certificate (signed by the student, the guide and the HOD) stating that the changes suggested by the examiners are incorporated in the revised report. Also a summary of the changes made in the revised report as per the suggestions of the examiners is to be submitted (as a separate manuscript) with the revised report. If the result of the second evaluation is worth E grade, (s)he will have to appear for the end semester examinations along with regular students. This provision is only applicable for Project Work evaluation
  .
- Failed or improvement candidates will have to appear for the end semester examinations along with regular students
  .

REVALUATION

Awarding of a higher grade after revaluation may be given only after a second revaluation.

IMPROVEMENT/SUPPLEMENTARY

maximum of two courses (Core or Elective) can be improved in each semester. Improvement of a particular semester can be done only once. The student shall avail the improvement chance in the succeeding year after the successful completion of the semester concerned. The internal marks already obtained will be carried forward to determine the grades/marks in the improvement examination. If the candidate fails to appear for the improvement examination after registration, or if there is no change in the results of the improvement examination appeared, the marks/grades obtained in the first appearance will be retained.

Improvement and supplementary examinations cannot be done simultaneously.

6 PATTERN OF QUESTION PAPERS

Duration of End Semester examinations for both theory and practical courses shall be 3 hours.

QUESTION PAPERS - THEORY

Section	No of Questions		Weightage for each question	Total
	To be Asked	To be Answered		
A: Short answer questions+	12	12	1	12
B: Short Essay	9	6	2	12
C: Essays*	6	3	4	12
			Total	**36**

+MCQ / fill in the blank / matching /one word / etc. Each question is to be answered in 7 minutes duration and should extract the critical knowledge acquired by the candidate in the subject.

Programs / Psuedocode / Problems / Derivations / Narrations. A question can have subdivisions. Each question is to be answered in 30 minutes. May be asked as a single question or in parts.

QUESTION PAPERS - PRACTICAL

Mark distribution for practical courses shall be as follows.

Component	Weightage
Algorithm/Flow diagram/UI diagram/Class Diagram	1
Implementation	1
Result/Output	1
Record	1
Viva	1
Total	**5**

PROJECT WORK

Total weightage for Project Work shall be 50.
Hence the total grade points shall be 200 (50 x 4).

The scheme of evaluation for project work shall be:

Component	Grade Points
Relevance of the Topic, Statement of Objectives, Methodology	15
Quality of Literature Survey/Product Review	15
Quality of Analysis Phase	10
Quality of Design Phase	10
Quality of Implementation/Simulation	10
Quality of Testing/Result Analysis	10
Quality of Contributions	10
Identification of Future Work	15
Quality of Project Report	50
Publications/Presentations out of the Project Work*	10
Quality of Presentation	15
Demonstration of the Project Work	15
General Viva Voce	15
Total	**200**
Grade is calculated by dividing total number of points obtained by a student by 50	

*In case at least one student of the batch has a publication/presentation out of his/her project work in a workshop/conference/journal/IT fest etc, this score is to be awarded for

the student; no other students will deserve score for this component! If none of the students in the batch could make such an edge, then the score for this component is to be added with the component "Identification of Future Work".

7 CREDIT SYSTEM

Each course shall have a specific credit (whole number) depending on the academic load and the nature and importance of the course. The credit associated with each course is as listed in the prescribed scheme and syllabi.

Direct Grading System based on a 5 point scale is used to evaluate the performance (External and Internal Examination of students).

- One Credit is equivalent to 4 periods of 60 minutes each, for theory and practical.
- Total credits of the PG Diploma in Data Science Programme shall be 40. The following is the semester wise credits a student must earn for the award of the degree:

Semester	Duration	Credits
I	Six Months	20
II	Six Months	20
Total	**12 Months**	**40**

8 DIRECT GRADING SYSTEM

- Direct Grading System based on a 4 point scale is used to evaluate the performance (external and internal examination of students).
- Each course is evaluated by assigning marks with a letter grade (A, B, C, D, E).

Grade	Performance	Grade Point	Grade Range
A	Excellent	4	3.50 - 4.00
B	Very Good	3	2.50 - 3.49
C	Good	2	1.50 - 2.49
D	Average	1	0.50 - 1.49
E	Poor	0	0.00 - 0.49

- Each course is evaluated by assigning a letter grade (A,B,C,D or E) to that course by the method of direct grading. The internal (weightage =1) and external weightage = 4) components of a course are separately graded and then combined to get the grade of the course after taking into account of their weightage.
- An aggregate of C grade (external and internal put together) is required in each course for a pass and also for awarding the degree.
- A student who fails to secure a minimum grade for a pass in a course will be permitted to write the examination along with the next batch.
- After the successful completion of a semester, Semester Grade Point Average (SGPA) of a student in that semester is calculated using the formula given below. For the successful completion of a semester, a student should pass all courses. However, a student is permitted to move to the next semester irrespective of SGPA obtained.

- SGPA of the student in that semester is calculated using the formula

$$\text{SGPA} = \frac{\text{Sum of the credit points of all courses in a semester}}{\text{Total credits in that semester}}$$

- The Cumulative Grade Point Average (CGPA) of the student is calculated at the end of a programme. The CGPA of a student determines the overall academic level of the student in a programme and is the criterion for ranking the students. CGPA can be calculated by the following

$$\text{CGPA} = \frac{\text{Total credit points obtained in two semesters}}{\text{Total credits acquired}}$$

- SGPA and CGPA shall be rounded off to two decimal places. CGPA determines the broad academic level of the student in a programme and is the index for ranking students (in terms of grade points).
- An overall letter grade (Cumulative Grade) for the entire programme shall be awarded to a student depending on her/his CGPA .

9 Grade Cards

The Institution shall issue to the students grade/marks card (by online) on completion of each semester, which shall contain the following information:

i. Name of the Institution.

ii. Title of the Programme – Post Graduate Diploma in Data Science.

iii. Semester concerned.

v. Name and Register Number of the student.

vi. Code number, Title and Credits of each course opted in the semester.

vii. Internal marks, External marks, total marks, Grade point (G) and Letter grade in each course in the semester.

viii. The total credits, total credit points and SGPA in the semester.

The Final Grade Card issued at the end of the final semester shall contain the details of all courses taken during the entire programme including those taken over and above the prescribed minimum credits for obtaining the degree. The Final Grade Card shall show the CGPA and the overall letter grade of a student for the entire programme.

10 Award Of Degree

The successful completion of all the courses prescribed for the PG Diploma in Data Science programme with C grade shall be the minimum requirement for the award of PG Diploma in Data Science programme degree.

11 Grievance Redressal Committee

College level

The College shall form a Grievance Redressal Committee in each department comprising of course teacher and one senior teacher as members and the HOD as Chairman. The Committee shall address all grievances relating to the internal assessment grades of the students. There shall be a college level Grievance Redressal Committee comprising of student advisor, two senior teachers and two staff council members (one shall be an elected member) as member and the Principal as the Chairperson.

12 Transistory Provision

Notwithstanding anything contained in these regulations, the Principal shall, for a period of one year from the date of coming into force of these regulations, have the power to provide by order that these regulations shall be applied to any programme with such modifications as may be necessary.

Programme Structure

LEGEND	
Item	**Description**
C	Credits
E	External Component (%)
I	Internal Component (%)
L	Lecture Hours
P	Practical Hours
T	Total

Course No	Course Code	Title	Hours		Marks		Credits
			T	P	Int.	Ext.	
Semester I							
1	VDST101	Fundamentals of Big data	5	0	20	80	4
2	VDST102	Data Mining Concepts	5	0	20	80	4
3	VDST103	Statistical techniques for data science	5	0	20	80	4
4	VDST104	R programming	4	2	20	80	4
5	VDSP101	Programming Laboratory: R Programming	0	4	20	80	4
			25		500		20
Semester II							
6	VDST201	Machine Learning	5	0	20	80	4
7	VDST202	Data Analysis using R programming	2	3	20	80	4
8	VDST203	Predictive analytics and data modeling	5	0	20	80	4
9	VDSP201	Project		10	50	150	8
			25		500		20

Semester I

Fundamentals of Big Data

Course Number: VDST101

Contact Hours per Week: 5T

Number of Credits: 4

Number of Contact Hours: 90

Course Evaluation: Internal – 20 Marks + External – 80 Marks

Learning outcomes

1. To cover the basics of big data.
2. To familiarize with big data technology and tools.

Course Outline

Unit I [18 T]

Introduction to Big Data : Definition & importance of Big Data ,Four dimensions of big data, Volume, Velocity, Variety, Veracity, Importance of big data, Structured data, Unstructured data,The role of a CMS in big data management, Integrating data types into a big data environment, Distributed computing and big data; Big data stack: layer 0,1 and 2; Big data management, Operational databases ,Relational databases ,Non relational databases, NoSQL, Key-value pair databases ,Document databases, Columnar databases, Graph databases ,Spatial databases.

Unit II [18 T]

Big Data analysis: Basic analytics, Operationalized analytics, Modifying business intelligence products to handle big data, Big data analytics examples, Analytics solutions, Text analytics, Exploring unstructured data, Understanding text analytics, Analysis and extraction techniques, The extracted information, Text analytics tools for Big Data, Custom applications for big data analysis, R environment, Google prediction API, Characteristics of a big data analysis framework.

Unit III [18 T]

NoSQL databases: Types, Advantages over relational databases; MongoDB: Introduction, MongoDB philosophy, The data model, Designing the database, Collections, Documents, Data types,The _id Field, Indexes, Viewing available databases and collections, Opening a database, Inserting data, Querying for data, Retrieving documents, Aggregation commands, Grouping results, Conditional operators, Specifying an array of matches, Applying criteria for

search , $slice , $size , $exists , $type , $elemMatch , $not (meta-operator), update() , save() , $inc, $set, $unset, $push, $pushAll,$addToSet, Removing elements from an array, Atomic operations, Modifying and returning a document atomically, Renaming a collection, Removing data, Referencing a database, Implementing index, Related functions , min() and max().

Unit IV [18 T]

Hadoop : History, Components, HDFS, MapReduce basics, Origins of MapReduce, Map function, Reduce function, Putting them together; Hadoop common components, Application development in Hadoop, Pig and Pig Latin, Load, Transform, Dump and store, Hive, Jaql, Getting our data into Hadoop, Basic copy data, Flume, Zookeeper, HBase ,Oozie,Lucene, Avro.

Unit V [18 T]

Understanding MapReduce: Key/value pairs, The hadoop java API for MapReduce, The mapper class, The reducer class, The driver class,Writing simple MapReduce programs, Hadoop-provided mapper and reducer implementations, Hadoop-specific data types, The writable and writablecomparable interfaces, Wrapper classes, Input/output - Inputformat and Recordreader,Outputformat and recordwriter; Implementing wordcount using streaming, Analyzing a large dataset, Summarizing the UFO data, Summarizing the shape data, A relational view on data with Hive, Creating a table for the UFO data, Iinserting the UFO data , Redefining the table with the correct column separator, Creating a table from an existing file, SQL views.

References

1. Hurwitz, Alan Nugent, Fern Halper and Marcia Kaufman, *Big Data for Dummies* , John Wiley & Sons, 2013

2. Eelco Plugge, Peter Membrey and Tim Hawkins ,*The Definitive Guide to MongoDB: The NOSQL Database for Cloud and Desktop Computing*, Apress, I Edition

3. Chris Elaton, Derk Deroos, Tom Deutsch, George Lapis and Pual Zikopoulos, *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*, McGraw Hill Professional, 2011

4. Garry Turkington, *Hadoop Beginner's Guide,* Packt Publishing Ltd, 2013

Data Mining Concepts

Course Number: VDST102

Contact Hours per Week: 5

Number of Credits: 4

Number of Contact Hours: 90

Course Evaluation: Internal – 20 Marks + External – 80 Marks

Learning Outcomes

- Introduce basic data mining concepts.
- Understand different algorithms and techniques
- To develop practical problem solving skills using data mining.

Course Outline

Module I [18T]

Data mining: Meaning, Definition, Goals, Scope, Related technologies, Stages involved in data mining, Data mining techniques , Major issues in data mining, Applications.

Module II [18T]

Data objects and attribute types attribute generalization and relevance, Class comparison, Statistical measures, Data visualization, Measures of similarity and dissimilarity; Data preprocessing: Overview, Data cleaning, Data integration, Data reduction, Data transformation, Discretization, Generating concept hierarchies.

Module III [18T]

Mining frequent patterns, associations and correlations: Basic concepts and methods, Frequent itemset mining methods -Apriori algorithm, Pattern growth approach, Framing association rules, Pattern evaluation methods, Pattern mining concepts, Mining in multi level and multidimensional space, Constraint based frequent pattern mining.

Module IV [18T]

Classification: Basic concepts, Decision tree induction, Nearest neighbor methods, Bayes classification methods; Cluster Analysis: Basic concepts and algorithms, Basic issues in clustering, Partitioning and hierarchical methods, Conceptual clustering, and density based methods, Advanced techniques, Evaluation of clustering.

Module V [18T]

Text mining ,Web mining, Spatial mining, Illustration of mining real data , Preprocessing data from a real domain , Applying various data mining techniques to create a comprehensive and accurate model of the data clustering.

References

1. Jiawei Han, Micheline Kamber and Jian Pei, *Data Mining Concepts and Techniques*, Morgan Kaufmann Publishers, 2011

2. Pang-Ning Tan,Michael Steinbach and Vipin Kumar , *Introduction to Data Mining*, Pearson Education Inc , 2003

4. Alex Berson and Stephen J. Smith, *Data Warehousing, Data Mining & OLAP*, *Computing* Mcgraw-Hill, Tata McGraw-Hill Education, 2004

5. K.P. Soman, Shyam Diwakar and V. Ajay, *Insight into Data mining Theory and Practice*, Prentice Hall of India, 1st Edition

6. G. K. Gupta, *Introduction to Data Mining with Case Studies*, PHI Learning Pvt. Ltd, 3rd Edition

Statistical Techniques for Data Science

Course Number: VDST103

Contact Hours per Week: 5

Number of Credits: 4

Number of Contact Hours: 90

Course Evaluation: Internal – 20 Marks + External – 80 Marks

Learning Outcomes

- Introduce basic concepts in statistics.
- Understand nature of data and different types of hypotheses

Course Outline

Module I [20T]

Probability Distributions: Definition classical, empirical, Addition and multiplication laws, Conditional probability, random variable, Probability functions, Discrete distribution-Binomial distribution, Poisson distribution. Continuous distribution- Normal distribution, lognormal distribution. Exponential Distribution (Concept and applications only).

Module II [20T]

Sampling Distributions: Theory of sampling distributions, Standard error, Sampling distribution of Sample mean, Chi square distribution, Student's t distribution, F distribution. (Concept and applications only)

Module III [10T]

Statistical Estimation : Point estimation, Properties of point estimation ,Unbiasedness, consistency, sufficiency, efficiency. Interval estimation, Confidence intervals.

Module IV [15T]

Testing of Hypothesis: Test of hypothesis- Null and alternative hypothesis, Type I and Type II errors, Critical region, Level of significance, *P Value*, Power of a test, Critical value, Neymann Pearson Lemma

Module V [25T]

Module 5: Parametric and Non parametric Tests: Parametric Tests- t, F, Z test. Non parametric Tests – Kolmogrov smirnov test, Kruskal Wallis test, Mann Whitney U test, Wilcoxon rank sum test, Welch's t test, Plots to check normality. Analysis of variance and covariance

References

1. S.C.Gupta and V.K.Kapoor (2010) *Fundamentals of Mathematical Statistics*, Eleventh thoroughly revised edition, Sultan Chand & Sons, New Delhi
2. Sheldon Ross(2014), *A first course in Probability* , Ninth edition, Pearson Education Inc.
3. E.L. Lehmann and Joseph P. Romano (2005), *Testing Statistical Hypothesis*, Springer.
4. George Casella and Roger L. Berger (2001), *Statistical Inference*, 2nd Edition

R programming

Course Number: VDST104
Contact Hours per Week: 4T+2P
Number of Credits: 4
Number of Contact Hours: 108
Course Evaluation: Internal – 20 Marks + External – 80 Marks
Learning outcomes

- **Understand the R programming environment**
- **Ability to manipulate data in R**
- **Visualize the data in an appropriate way**
- **Use statistical tests in R**

Course Outline

Module I [12T+6P]

Introduction to R: R features, R script, R data types- vectors, lists, matrices, arrays, factors, data frames, R variables, R operators, R packages, I/O functions, Control Structures- Decision Making,- if, if-else, switch, Loops- Repeat, While, For, Statements- break, next

Module II [14T+8P]

Function in R: Built in functions to manipulate vectors, lists, matrices, arrays, factors, data frames, User defined functions, data reshaping

Module III [16T + 8P]

Data Manipulation: Import /Export of excel, binary, csv, xml, database files, Data manipulation using plyr and dplyr libraries

Module IV [16T + 8P]

Data Visualization: Pie charts, Bar charts, Box plots, Histograms, Line graphs, Scatter plots, Visualization packages for univariate, bivariate and multivariate

Module V [14T + 6P]

Statistical Inference using R: Descriptive statistics functions, Hypothesis testing, Type –I and Type –II errors, students t-test, Welsch t-test, Wilcoxon Rank sum test, ANOVA, ANCOVA, F-test, Z-test, normality functions.

References

1. Jaynal Abedin, Kishore Kumar R, *Data Manipulation with R*, Packt Publishing Ltd, 2015
2. Eric Mayor, *Learning Predictive Analytics with R*, Packt Publishing Ltd, 2015
3. Sudha G Purohit, Sharad D. Gore, Shailaja R Desmukh , *Statistics using R*

Programming Laboratory: R Programming

Course Number: VDSP101
Contact Hours per Week: 0T+4P
Number of Credits: 4
Number of Contact Hours: 72
Course Evaluation: Internal – 20 Marks + External – 80 Marks
Learning outcomes

- **Practical knowledge of R**

Course Outline

1. Demonstrate various R data types
2. Demonstrate if-else in R
3. Demonstrate Switch in R
4. Demonstrate loops in R
5. Demonstrate built in function to manipulate various data types
6. Demonstrate user defined functions
7. Demonstrating import and export of various databases
8. Demonstrate usage of plyr libraries
9. Demonstrate usage of dplyr libraries
10. Demonstrate univariate visualization packages
11. Demonstrate multivariate visualization packages
12. Demonstrate descriptive statistics in R
13. Demonstrate t-test in R
14. Demonstrate Welsch t-test in R
15. Demonstrate Wilcoxon t-test in R
16. Demonstrate F-test
17. Demonstrate Z-test
18. Demonstrate normality functions
19. Demonstrate ANOVA
20. Demonstrate ANCOVA

SEMESTER 2

Machine learning

Course Number: VDST201
Contact Hours per Week: 5T
Number of Credits: 4
Number of Contact Hours: 90
Course Evaluation: Internal – 20 Marks + External – 80 Marks

Learning outcomes

- To develop an appreciation for what is involved in learning from data.
- To understand a wide variety of learning algorithms.
- To understand how to apply a variety of learning algorithms to data.
- To understand how to perform evaluation of learning algorithms and model selection.

Course Outline

Module I [18T]

Techniques of Machine Learning: Supervised learning, Unsupervised learning, Semi supervised learning, Reinforcement learning, Machine Learning Algorithm, Application of machine learning.

Module II [18T]

Feature Engineering: Feature Improvement: Dealing with missing data, Standardization & normalization; Feature Selection: Statistics based feature selection, model based feature selection; Feature transformation: Principal Component Analysis, linear discriminant analysis; Feature Learning.

Module III [18T]

Regression: Linear regression- univariate, multi variate models, Logistic Regression, Discriminant Analysis, Poisson Regression, Regularization methods- ridge, lasso, gradient descent.

Module IV [18T]

Clustering & Classification: K-nearest neighbors, Support Vector Machine, Kernel SVM, Random forest classifier, maximum entropy classifier, Artificial neural networks: Perceptrons, Single layer feed forward networks, Back propagation; Fuzzy Clustering.

Module V [18T]

Introduction to Deep Learning: Meaning and importance of deep learning, applications, Deep neural network, Convolutional neural network, Recurrent neural network, LSTM, Tensor Flow.

References:

1. Rogers S and Girolami M, *A first course in Machine Learning*, CRC Press, 2011
2. Bishop C, *Pattern Recognition and Machine Learning*, Springer 2007
3. Cory Lesmeister, *Mastering Machine Learning with R*, Packt, 2nd Edition
4. Mitchell T, *Machine Learning*, McGraw-Hill, 1997
5. Barber D, *Bayesian Reasoning and Machine Learning*, Cambridge University Press, 2012
6. Duda, Hart and Stork, *Pattern Classification*, Wiley-interscience,2012
7. Sinan Ozdemir, Divya Susarla, *Feature Engineering Made Easy: Identify unique features from your dataset in order to build powerful machine learning systems*, Packt Publishing Ltd, 2018

Data Analysis Using R

Course Number: VDST202

Contact Hours per Week: 3T+3P

Number of Credits: 4

Number of Contact Hours: 108

Course Evaluation: Internal – 20 Marks + External – 80 Marks

Learning outcomes

- **Ability to choose the appropriate test to analyze data**
- **Learn to analyze data and interpret the analytical results**

Course Outline

Module I [10T+8P]

Mining & Classification in R: Algorithms: Apriori, FP growth, DT, ID3, C4.5, CART, Bayes classification, Rule based classification, case studies

Module II [8T+10P]

Advanced Classification Techniques: Ensemble methods, Bayesian Belief networks, k-nearest, SVM, Back propagation, a case study

Module III [12T+12P]

Clustering Techniques and Outlier Detection: k-means, k-medoids, CLARANS, STING, outlier detection methods, a case study

Module IV [12T+12P]

Regression: Linear, Multiple, logistic, regularization, analysis of variance and covariance, case studies

Module V [12T+12P]

Advanced analytical models: Time series analysis - ARIMA models, ARMA models, Text analysis - Term frequency inverse document frequency, case studies

References:

1. Michael J Crawley ,*The R Book*, John Wiley & Sons, 2012
2. Bater Makhabel, *Learn data mining with R*, Packt Publishing Ltd, 2015
3. EMC Education Services, *Data Science and Big Data Analytics- Discovering, Analyzing, Visualizing and Presenting Data*, John Wiley & Sons, 2015
4. Eric Mayor,*Learning Predictive Analytics with R*, Packt Publishing Ltd, 2015

Predictive Analytics And Data Modeling

Course Number: VDST203

Contact Hours per Week: 5T

Number of Credits: 4

Number of Contact Hours: 90

Course Evaluation: Internal – 20 Marks + External – 80 Marks

Learning outcomes

- Represent and manipulate data in effective ways.
- Manipulate data using packages/tools and by ad hoc data handling.
- Use mathematical, computational and statistical tools to detect patterns and model performance.

Course Outline

Module I [10T]

Correlation Techniques: Data types or levels of measurement- Nominal,ordinal, interval and ratio, Measures of Correlation, Simple correlation, Partial correlation and Multiple correlation.

Module II [10T]

Regression Analysis: Simple linear regression, Gauss Markov theorem, Basics of fitting and residual analysis, Multiple linear regression, model adequacy techniques, Path Analysis (concept only).

Module III [20 T]

Logistic regression: Moving from linear to logistic regression, Model assumptions and Odds ratio, ROC curve and KS statistic. (Concept and Applications using real data sets)

Module IV [30 T]

Time series analysis: Components of time-series, additive and multiplicative models ,Methods for measurement of trends, Methods for measurement of seasonal fluctuations, forecasting, Autocorrelation, ARIMA Model, ARMA Model.(Concept and Applications using real data sets)

Module V [20 T]

Multivariate Data Analysis: Multivariate data, plotting multivariate data, matrix scatter plot, Cluster analysis, Principal Component analysis, Discriminant analysis, Factor analysis.(Concept and Applications using real data sets)

References

1. D.C.Montgomery, E.A.Peck and G.G.Vining (2006) *Introduction to linear regression analysis*, Third Edition, Wiley India Private Ltd, New Delhi.
2. S.C.Gupta and V.K.Kapoor (2014) *Fundamentals of Applied Statistics*, Fourth thoroughly revised edition, Sultan Chand & Sons, New Delhi.
3. P.Mukhopadhyay (2016) *Applied Statistics*, Second Edition (thoroughly revised), Books and Allied (P) Ltd., Kolkata.
4. S.C.Gupta and V.K.Kapoor (2010) *Fundamentals of Mathematical Statistics*, Eleventh thoroughly revised edition, Sultan Chand & Sons, New Delhi.
5. H¨ardle, W. and Simar, L. (2003). Applied Multivariate Statistical Analysis